Econometria
Propiedades Estadísticas Modelo de Regresión Lineal Simple

Pasquini, Ricardo

IAE
Universidad Austral

27 de agosto de 2025

Modelo de regresión simple

Repaso: Estimadores OLS

$$\underbrace{Y_i}_{\text{Variable a explicar o Dependiente}} = \underbrace{\beta_0}_{\text{Constante o intercepto}} + \beta_1 \cdot \underbrace{X_i}_{\text{Variable explicativa 1 (independiente)}} + \underbrace{\varepsilon_i}_{\text{Error}}$$

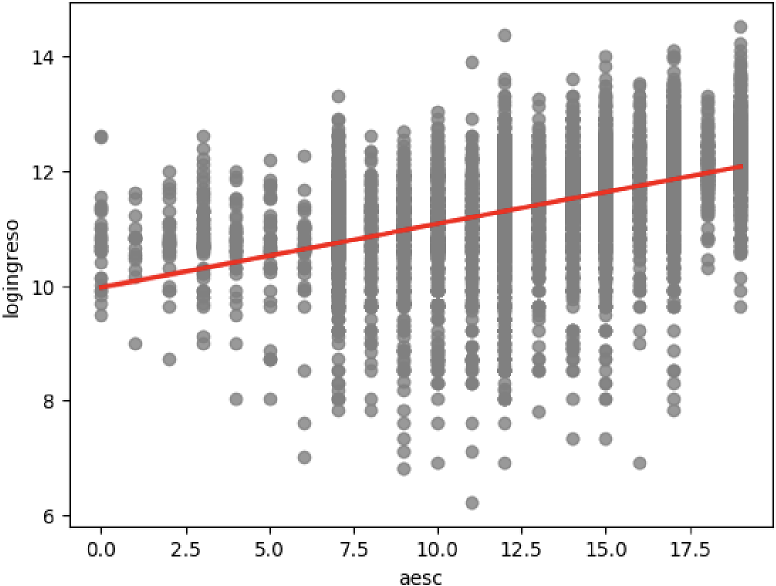
- $\hat{\beta}_0, \hat{\beta}_1$ surgen de la minimización de la suma de los residuos al cuadrado (OLS).
- Los estimadores OLS son:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

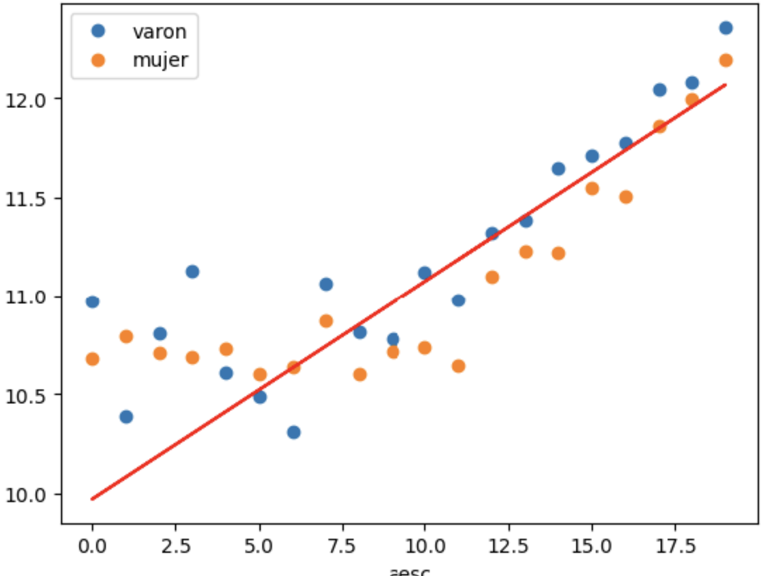
Estimadores de Mínimos Cuadrados Ordinarios (OLS)

Ejemplo: Ingresos y Años de Escolaridad



Estimadores de Mínimos Cuadrados Ordinarios (OLS)

Ejemplo: Regresión y CEF



Propiedades Estadísticas de los Coeficientes

Insesgadez

- **Ausencia de Sesgo (Insesgadez):** Esta propiedad establece que, *en valor esperado*, los coeficientes estimados serán iguales a los verdaderos coeficientes poblacionales.
- Formalmente, se define como:

$$E(\hat{\beta}_j) = \beta_j$$

- Intuición: Aunque las estimaciones individuales pueden variar debido a la aleatoriedad inherente del muestreo, si pudiéramos repetir esta estimación en múltiples muestras, el promedio de estas estimaciones coincidiría con el valor verdadero.

Propiedades Estadísticas de los Coeficientes

Insesgadez

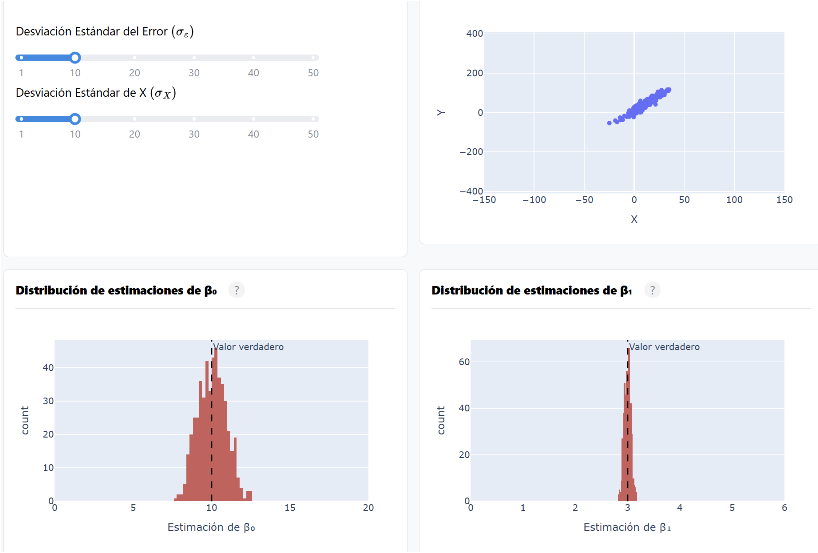


Figura: Simulación. Ver en SimuEcon.com

Idea de la Simulación

- **Objetivo**: Mostrar que los coeficientes del modelo OLS son insesgados.
- **Pasos de la simulación**:
  - **Definición de parámetros**: Establecemos los valores verdaderos de los coeficientes asumidos verdaderos (β_0 y β_1), tamaño de muestra y número de simulaciones.
  - **Iteración**: Realizar múltiples simulaciones.
    - Generamos pares (x,y) provenientes de la poblacion (cumplen $y = \beta_0 + \beta_1 X + \epsilon$)
    - Ajustamos el modelo de regresión lineal y obtenemos los estimadores de los coeficientes.
    - Almacenar los valores estimados de los coeficientes.
  - **Visualización**: Construimos histogramas de los valores estimados de los coeficientes.
  - **Estadísticas descriptivas**: Calculamos media y desviación estándar de los valores estimados de los coeficientes.

Supuestos de OLS para garantizar que no hay sesgo

Detalles Técnicos

- Supuestos necesarios para la ausencia de sesgo del estimador OLS:
  1. Linealidad en los parámetros (el modelo poblacional es $y = \beta_0 + \beta_1 X + \varepsilon$)
  2. Muestreo aleatorio (los valores (x_i, y_i) son variables aleatorias del modelo poblacional)
  3. La esperanza condicional del error es cero ($E(\varepsilon|X) = 0$). En este caso, garantizado por OLS.
- Demostración:
  - Queremos demostrar que $E(\hat{\beta}_1) = \beta_1$
  - Partimos de la definición de $\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})\varepsilon_i}{SST_x}$
  - Utilizamos que los datos provienen de la población, por lo que $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
  - Luego $E(\hat{\beta}_1|X) = E(\beta_1 + \frac{\sum(x_i - \bar{x})\varepsilon_i}{SST_x}|X)$
  - $E(\hat{\beta}_1|X) = \beta_1 + \frac{1}{SST_x}\sum(x_i - \bar{x})E(\varepsilon_i|X)$
  - Como $E(\varepsilon_i|X) = 0$ (supuesto 3), entonces $E(\hat{\beta}_1|X) = \beta_1$
  - Luego $E(\hat{\beta}_1) = \beta_1$

Propiedades Estadísticas de los Coeficientes

Varianza

- Aunque insesgadas, nuestras estimaciones siempre exhibirán cierto grado de **varianza**, que cuantifica la incertidumbre alrededor del coeficiente estimado.
- Formalmente, se define como:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2}$$

- σ^2 es la varianza de los residuos.
- $\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2$ es la suma de los cuadrados de las desviaciones de la variable independiente X_j con respecto a su media.
- Con múltiples variables se incorporará un factor de corrección adicional que dependerá de la correlación entre las variables independientes. (próxima clase)

Propiedades Estadísticas de los Coeficientes

Varianza

- La teoria indica que la varianza de nuestras estimaciones está influenciada por dos factores clave:
  - **Error en el modelo:** La presencia de variación no explicada en la variable dependiente (Y) contribuye a la varianza de nuestras estimaciones. Este error puede atribuirse a factores no incluidos en el modelo o a la aleatoriedad inherente en los datos.
  - **Variabilidad de la variable independiente:** Una mayor dispersión en los valores de nuestra variable independiente (X) conduce a una menor varianza en nuestras estimaciones de coeficientes. Esto se debe a que un rango más amplio de valores de X proporciona más información para estimar la relación con Y.

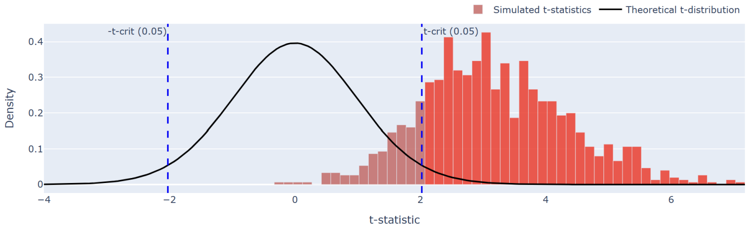
Varianza del Estimador OLS - Derivación

- Para derivar este resultado partimos de $\hat{\beta}_1 = \beta_1 + \frac{\sum(x_i - \bar{x})\varepsilon_i}{SST_x}$
- Luego $Var[\hat{\beta}_1|x] = \frac{1}{SST_x^2}\sum(x_i - \bar{x})^2 Var(\varepsilon_i)$

- Bajo ciertos supuestos, los coeficientes siguen una distribución normal:
$$\hat{\beta} \sim N(\beta, Var(\beta))$$
- El valor estandarizado sigue una Normal Estándar:
$$Z = \frac{\hat{\beta} - \beta}{\sqrt{Var(\beta)}} \sim N(0, 1)$$
- En la práctica, usamos:
$$T = \frac{\hat{\beta} - \beta}{\sqrt{\hat{Var}(\beta)}} \sim T_{n-1}$$

Test de Hipótesis

Distribución T y Potencia ?



Nivel de Significancia (α)	Potencia (%)
α = 0.01	62.50%
α = 0.05	85.10%
α = 0.10	91.40%

Figura: Distribución del estadístico T bajo H_0

Test de Hipótesis

- Test típico:

$$\begin{cases} H_0 : \beta = 0 \\ H_a : \beta \neq 0 \end{cases}$$

- Interpretación: Si H_0 es válida, X no tiene efecto sobre Y
- **Lógica del Test:**
  1. Asumimos $\beta = 0$ (Hipótesis nula)
  2. Derivamos la distribución de probabilidad según esa hipótesis
  3. Observamos nuestra estimación $\hat{\beta}$ y construímos el estadístico de prueba $\hat{T}$
  4. Si $\hat{T}$ es muy atípico, rechazamos H_0
- Cuán atípico es $\hat{T}$? **P-valor:** Probabilidad de observar un valor tan extremo como $\hat{T}$ bajo H_0

Test de Hipótesis

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                ingreso   R-squared:                       0.176
Model:                            OLS   Adj. R-squared:                  0.176
Method:                 Least Squares   F-statistic:                     1609.
Date:                Wed, 27 Aug 2025   Prob (F-statistic):          4.94e-324
Time:                        11:57:44   Log-Likelihood:            -1.1779e+05
No. Observations:                9062   AIC:                         2.356e+05
Df Residuals:                    9060   BIC:                         2.356e+05
Df Model:                           1
Covariance Type:                  HC0
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -4.489e+04   3867.026    -11.607      0.000   -5.25e+04   -3.73e+04
aesc        1.325e+04    330.296     40.115      0.000    1.26e+04    1.39e+04
==============================================================================
Omnibus:                     7371.512   Durbin-Watson:                   1.790
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           361472.985
Skew:                           3.555   Prob(JB):                         0.00
Kurtosis:                      33.113   Cond. No.                         50.6
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
```

Figura: Interpretación del P-valor. Ver en SimuEcon.com

Test de Hipótesis

Ejemplo Pay-for-Performance - Lazear 2000

VOL. 90 NO. 5 LAZEAR: PERFORMANCE PAY AND PRODUCTIVITY 1353

TABLE 3—REGRESSION RESULTS

Regression number	Dummy for PPP person-month observation	Tenure	Time since PPP	New regime	R^2	Description
1	0.368 (0.013)				0.04	Dummies for month and year included
2	0.197 (0.009)				0.73	Dummies for month and year; worker-specific dummies included (2,755 individual workers)
3	0.313 (0.014)	0.343 (0.017)	0.107 (0.024)		0.05	Dummies for month and year included
4	0.202 (0.009)	0.224 (0.058)	0.273 (0.018)		0.76	Dummies for month and year; worker-specific dummies included (2,755 individual workers)
5	0.309 (0.014)	0.424 (0.019)	0.130 (0.024)	0.243 (0.025)	0.06	Dummies for month and year included

Notes: Standard errors are reported in parentheses below the coefficients.
Dependent variable: ln output-per-worker-per-day.
Number of observations: 29,837.

Figura: Lazear 2000

Takeaways

- Reducir el error del modelo ayuda a mejorar la precisión de los coeficientes individuales
- Buscar ampliar la varianza de las variables explicativas incrementa la precisión.
- El test de hipótesis se realiza bajo una distribución de probabilidades centrada en la hipótesis nula (típicamente de 0 efecto). La rechazamos si encontramos un valor atípicamente alto (bajo). El p-valor, es la medida de cuán atípico es el valor encontrado.

Takeaways

- Puesto que solo podemos testear usando el estadístico T, y este estadístico considera el efecto en relación a su standard error, la *significatividad estadística* solo habla de un concepto relativo: cuan grande o chico es un efecto en relación a la precisión con la que fue estimado.